TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Remarques et perspectives sur les langages de prégroupe d'ordre 1/2

Denis Béchet, Annie Foret
INRIA & Université de Rennes 1, IRISA
Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 Rennes Cedex
France
{Denis.Bechet,Annie.Foret}@irisa.fr

Mots-clefs – Keywords

Acquisition automatique, inférence grammaticale, modèle de Gold, prégroupe.
Automatic acquisition, grammatical inference, model of Gold, pregoup.

Résumé - Abstract

Cet article traite de l'acquisition automatique des grammaires de Lambek, utilisées pour la modélisation syntaxique des langues. Récemment, des algorithmes ont été proposés dans le modèle d'apprentissage de Gold, pour certaines classes de grammaires catégorielles. En revenche, les grammaires de Lambek rigides ou k-valuées ne sont pas apprenables à partir des chaînes. Nous nous intéressons ici au cas des grammaires de prégroupe. Nous montrons que la classe des grammaires de prégroupe n'est pas apprenable à partir des chaînes, même si on limite fortement l'ordre des types (ordre 1/2) ; notre preuve revient à construire un point limite pour cette classe.

The article is concerned by the automatic acquistion of some grammars introduced by Lambek that are used for modeling the syntax of natural languages. Recently, some algorithms have been proposed in the model of Gold for several classes of categorial grammars. On the other hand, rigid or k-valued Lambek calculus grammars are not learnable from strings. We study here pregroup grammars. We prove that rigid grammars are not learnable from strings even if the order of types are bound by a constant (1/2) ; Our proof gives a limit point for this class of grammars.

1 Introduction

Les grammaires de prégroupe ont été introduites récemment dans (Lambek, 1999) dans le domaine du traitement automatique des langues, comme une alternative aux autres modèles de la structure syntaxique des langues. Ce sont des grammaires *lexicalisées,* tout comme les grammaires catégorielles en général auxquelles les grammaires de prégroupe sont apparentées ; c'est-à-dire qu'elles associent des *types* à chaque mot du lexique, tandis que les règles avec lesquelles elles fonctionnent sont fixées ; ces règles, que ce soit dans les prégroupes ou dans les grammaires catégorielles de Lambek introduites auparavant, sont de nature logique et algébrique ; analyser une phrase revient alors à faire une déduction logique ou algébrique.

Nous nous intéressons plus particulièrement au problème de *l'acquisition automatique* de telles grammaires ; dans ce cadre, il s'agit d'obtenir les types associés aux mots, à partir d'exemples ; en se plaçant dans le modèle de Gold d'apprentissage, le processus est supposé converger lorsque le nombre d'exemples positifs est suffisamment grand. Ce paradigme a suscité des travaux récents, depuis les résultats encourageants de (Kanazawa, 1998) et concerne plusieurs équipes, notamment en France, citons les travaux tels que (Bonato & Retoré, september 2001; Dudau-Sofronie, décembre 2001).

L'interêt des prégroupes paraît multiple : ils offrent un moyen d'expression de constructions linguistiques complexes, par exemple (Bargelli & Lambek, 2001; Casadio & Lambek, 2001) montrent comment traiter des clitiques ; d'un autre point de vue, ils disposent d'un algorithme polynomial pour tester l'appartenance au langage ; enfin dans une perspective d'acquisition automatisée, être lexicalisé est un avantage : cela permet des mises à jour incrémentales lorsque des mots ou des usages nouveaux doivent être pris en compte.

L'article est organisé comme suit. La section 2 présente des exemples, les définitions nécessaires et des résultats connus. La section 3 donne ensuite la construction de point limite et la preuve pour les grammaires de prégroupe d'ordre 1/2. La section 4 conclut.

2 Exemples, définitions et prérequis

2.1 Prégroupes

Un *prégroupe* est un monoïde partiellement ordonné dans lequel tout élément a possède un adjoint à droite a^r et un adjoint à gauche a^l tel que $a^l \cdot a \leq 1 \leq a \cdot a^l$ et $a \cdot a^r \leq 1 \leq a^r \cdot a$
Ceci permet de parler d'adjoints itérés : $a^{(i-1)} = \left(a^{(i)}\right)^l,\ a^{(i+1)} = \left(a^{(i)}\right)^r,\ a^{(0)} = a$

Prégroupes pour la linguistique. Nous reprenons ici un exemple linguistique dû à Lambek (Lambek, 2001). Ces exemples utilisent les types de base suivants :

π_2	= deuxième personne	s_1	= énoncé au temps présent
p_1	= participe présent	p_2	= participe passé
o	= objet	q	= question en oui ou non
q'	= question		

– Voici un exemple de type $\leq s_1$:

You	have	been	seeing	her
π_2	$(\pi_2^r s_1 p_2^l)$	$(p_2 p_1^l)$	$(p_1 o^l)$	o

Les réductions successives portent ici sur des parties de types provenant de mots consécutifs : $\pi_2\pi_2^r \leq 1$, $p_2^l p_2 \leq 1$, $p_1^l p_1 \leq 1$ et $o^l o \leq 1$.

– Un exemple de type $\leq q$ (et aussi de type $\leq q'$ si $q \leq q'$) :

have	you	seen	her ?
$(qp_2^l\pi_2^l)$	π_2	(p_2o^l)	o

Ici les réductions peuvent se produire à un certain stade entre mots distants (p_2^l et p_2).

– un exemple de type $\leq q'$ (en reprenant des types ci-dessus — le tiret indique ici une trace) :

whom	have	you	seen – ?
$(q'o^{ll}q^l)$	$(qp_2^l\pi_2^l)$	π_2	(p_2o^l)

Prégroupes libres. En fait, dans les grammaires de prégroupe pour la linguistique, nous utilisons des prégroupes libres qui sont obtenus à partir du monoïde libre sur un alphabet $P^{(\mathbb{Z})}$ (l'ensemble des mots sur cet alphabet). Nous notons de façon générale X^* l'ensemble des mots (comprenant le mot vide) sur un alphabet X et X^+ l'ensemble des mots non vides sur l'alphabet X. Par la suite, Σ désigne un alphabet donné (les mots d'une langue naturelle).

Soit $(P, \leq)$ un ensemble ordonné fini de *types primitifs*. L'ensemble des *types de base* est $P^{(\mathbb{Z})} = \{p^{(i)} \mid p \in P, i \in \mathbb{Z}\}$ et l'ensemble des *types* $T_{(P,\leq)} = \left(P^{(\mathbb{Z})}\right)^* = \{p_1^{(i_1)} \cdots p_n^{(i_n)} \mid 0 \leq k \leq n, p_k \in P \text{ et } i_k \in \mathbb{Z}\}$ (l'ensemble des mots de longueur finie sur l'alphabet $P^{(\mathbb{Z})}$). Pour X et $Y \in T_{(P,\leq)}$, on dit que $X \leq Y$ si et seulement cette relation est déductible dans le système d'inférence suivant où $p, q \in P$, $n, k \in \mathbb{Z}$ et $X, Y, Z \in T_{(P,\leq)}$:

$$\frac{}{X \leq X}\,(ID) \qquad \frac{XY \leq Z}{Xp^{(n)}p^{(n+1)}Y \leq Z}\,(A_G) \qquad \frac{X \leq YZ}{X \leq Yp^{(n+1)}p^{(n)}Z}\,(A_D)$$

$$\frac{X \leq Y \quad Y \leq Z}{X \leq Z}\,(CUT) \qquad \frac{Xp^{(k)}Y \leq Z}{Xq^{(k)}Y \leq Z}\,(IND_G) \qquad \frac{X \leq Yp^{(k)}Z}{X \leq Yq^{(k)}Z}\,(IND_D)$$

$q \leq p$ si k est pair, et $p \leq q$ si k est impair

Cette construction, due à Buskowski, définit un prégroupe qui étend l'ordre $\leq$ sur les types primitifs P à $T_{(P,\leq)}$. Nous appelons *prégroupe libre simple* le prégroupe libre basé sur l'égalité des types primitifs. La règle de coupure peut être éliminée (Buszkowski, 2001).

Grammaire/langage de prégroupe libre, rigide, k-valuée.

– Une *grammaire du prégroupe libre basé sur* $(P, \leq)$ est un triplet $G = (\Sigma, I, s)$ tel que :
  – Σ est un alphabet fini ;
  – $I : \Sigma \mapsto \mathcal{P}^f(T_{(P,\leq)})$ assigne un ensemble fini de *types* à chaque élément de Σ ;
  – $s \in P$ est le *type principal* associé aux phrases correctes.
– Les grammaires associant au plus k types à chaque symbole de l'alphabet sont appelées grammaires k-valuées ou grammaires rigides si $k = 1$.
– *La grammaire assigne le type X à un mot $v_1 \cdots v_n$ de* Σ^* si et seulement si pour $1 \leq i \leq n$, $\exists X_i \in I(v_i)$ tels que $X_1 \cdots X_n \leq X$ dans le prégroupe libre.
– *Le langage* de cette grammaire noté $\mathcal{L}(G)$ est l'ensemble des mots de Σ^* dont on peut assigner le type s.

2.2 Apprentissage et points limites

Point limite. Une classe de langages a un *point limite* L ssi il existe une suite infinie $(L_n)_{n\in\mathbb{N}}$ de langages telle que : $\begin{cases} L_0 \subsetneq L_1 \cdots \subsetneq \cdots \subsetneq L_n \subsetneq \cdots \\ L = \bigcup_{n\in\mathbb{N}} L_n \end{cases}$

Propriété. Une classe de langages qui possède un point limite n'est pas apprenable.

3 Grammaires de prégroupe d'ordre $n/2$

La classe des grammaires de prégroupe k-valuées n'étant pas apprenable (Foret, 2002), nous nous sommes intéressés à des sous-classes avec l'espoir d'obtenir des résultats positifs. Nous pouvons notamment hiérarchiser les grammaires en bornant l'ordre des types.

Grammaire de prégroupe d'ordre $n/2$
- Une grammaire de prégroupe sur $(P, \leq)$ d'ordre $n \in \mathbb{N}$ est une grammaire où les types de base sont dans $P^{(-n\cdots n)} = \{a^{(i)} | a \in P \; et \; -n \leq i \leq n\}$.
- Une grammaire de prégroupe sur $(P, \leq)$ d'ordre $n + 1/2, n \in \mathbb{N}$ est une grammaire où les types de base sont dans $P^{(-n-1\cdots n)} = \{a^{(i)} | a \in P \; et \; -n-1 \leq i \leq n\}$.

L'utilisation des $1/2$ ordres permet de distinguer les grammaires avec des types symétriques autour des types primitifs (les grammaires d'ordre n) et les grammaires sans type central (les grammaires d'ordre $n + 1/2$).

Nous montrons dans cette section que la plus petite classe, à part la classe des grammaires d'ordre 0 qui n'est pas très intéressante, admet un point limite. Ainsi, toutes ces familles, sauf à l'ordre 0, ne sont pas apprenables.

Définitions de $\mathcal{R}$ et $\mathcal{R}^*$
$\mathcal{R} : Xp^l pY \xrightarrow{\mathcal{R}} XY$, $p \in P$, $X, Y \in (P \cup P^l)^*$.
$\mathcal{R}^*$: la fermeture transitive et symétrique de $\mathcal{R}$ (notation $\xrightarrow{\mathcal{R}^*}$).

Lemme. Pour $X \in (P \cup P^l)^*$ et $s \in P$, $X \leq s$ si et seulement si $X \xrightarrow{\mathcal{R}^*} s$

Preuve. $X \leq s$ ssi $X \leq s$ dans le système de déduction des prégroupes libres sans coupure. Or, dans ce système puisque l'ordre $\leq$ pour les types primitifs est l'identité (le prégroupe est simple), les règles (IND_G) et (IND_D) ne sont pas utilisables. De même, s étant un type primitif, la règle (A_D) est impossible. Donc, une déduction sans coupure ne comporte que l'axiome $s \leq s$ suivi éventuellement d'une ou plusieurs applications de la règle (A_G) qui correspond à la règle de réécriture $\xrightarrow{\mathcal{R}}$ appliquée à la partie gauche de l'inégalité

Lemme. $\mathcal{R}$ est fortement confluent et nœthérien sur $(P \cup P^l)^*$.

Preuve. Pour la confluence, il suffit de voir que lorsque la règle $\xrightarrow{\mathcal{R}}$ est applicable à deux endroits différents sur un même mot, les deux motifs ne peuvent pas se recouper : $XYb^l bZ \xleftarrow{\mathcal{R}} Xa^l aYb^l bZ \xrightarrow{\mathcal{R}} Xa^l aYZ$. Donc, dans les deux cas nous pouvons réduire les deux mots à XYZ : $XYb^l bZ \xrightarrow{\mathcal{R}} XYZ \xleftarrow{\mathcal{R}} Xa^l aYZ$.

Le système est nœthérien car toute application de la règle diminue la longueur du mot de deux lettres. Il ne peut pas y avoir de suite infinie de réécritures.

Lemme. Pour $X, Y \in (P \cup P^l)^*$ et $p, q, s \in P$, si $Xp^l qY \xrightarrow{\mathcal{R}} s$ alors $p = q$

Preuve. Dans un mot, les lettres disparaissent deux à deux. Une lettre de P^l va disparaître avec une lettre de P qui se trouve à sa droite et symétriquement une lettre de P avec une lettre de P^l à sa gauche. Donc, un couple de deux lettres consécutives p^l et q ne peut que disparaître au même moment lors de l'application de la règle $\xrightarrow{\mathcal{R}}$ et $p = q$.

Construction du point limite. Soient $P = \{p, q, r, s\}$ et $\Sigma = \{a, b, c, d, e\}$.

– Pour $n \geq 0$, soient $G_n = (\Sigma, \left\{ \begin{array}{l} a \mapsto (p^l)^n q^l \\ b \mapsto qpq^l \\ c \mapsto qr^l \\ d \mapsto rp^l r^l \\ e \mapsto rp^n s \end{array} \right\}, s)$ et $L_n = \mathcal{L}_{(P,=)}(G_n)$.

– Soient $G_* = (\Sigma, \left\{ \begin{array}{l} a \mapsto q^l \\ b \mapsto qp^l q^l \\ c \mapsto qr^l \\ d \mapsto rpr^l \\ e \mapsto rs \end{array} \right\}, s)$ et $L_* = \mathcal{L}_{(P,=)}(G_*)$.

Lemme. Pour $n \geq 0$, $L_n = \{ab^k cd^k e \mid 0 \leq k \leq n\}$.

Preuve. Les lemmes précédents nous indiquent que seuls les mots de ab^*cd^*e peuvent appartenir à L_n. Pour $i \geq 0$ et $j \geq 0$, $ab^i cd^j e \in ab^*cd^*e$ et nous avons :

$$\begin{array}{ll} ab^i cd^j e \in L_n & \iff \\ (p^l)^n q^l (qpq^l)^i qr^l (rp^l r^l)^j rp^n s \xrightarrow{\mathcal{R}*} s & \iff \quad (\mathcal{R}* \text{ confluent et nœthérien}) \\ (p^l)^n p^i (p^l)^j p^n s \xrightarrow{\mathcal{R}*} s & \end{array}$$

Quatre cas doivent être analysés.

– Si $i \leq n$ et $j \leq n$:
$$\begin{array}{ll} ab^i cd^j e \in L_n & \iff \\ (p^l)^{n-i} p^{n-j} s \xrightarrow{\mathcal{R}*} s & \iff \\ i = j & \end{array}$$

– Si $i \leq n$ et $j > n$: $ab^i cd^j e \in L_n \iff (p^l)^{n-i}(p^l)^{j-n} s \xrightarrow{\mathcal{R}*} s$ impossible !

– Si $i > n$ et $j \leq n$: $ab^i cd^j e \in L_n \iff p^{i-n} p^{n-j} s \xrightarrow{\mathcal{R}*} s$ impossible !

– Si $i > n$ et $j > n$: $ab^i cd^j e \in L_n \iff p^{i-n}(p^l)^{j-n} s \xrightarrow{\mathcal{R}*} s$ impossible !

Lemme. $L_* = \{ab^k cd^k e \mid k \geq 0\}$.

Preuve. Comme pour le lemme précédent, seuls les mots de ab^*cd^*e peuvent appartenir à L_n. Pour $i \geq 0$ et $j \geq 0$, $ab^i cd^j e \in ab^*cd^*e$ et nous avons :

$$\begin{array}{ll} ab^i cd^j e \in L_* & \iff \\ q^l (qp^l q^l)^i qr^l (rpr^l)^j rs \xrightarrow{\mathcal{R}*} s & \iff \\ (p^l)^i p^j p^n s \xrightarrow{\mathcal{R}*} s & \iff \\ i = j & \iff \\ ab^i cd^j e \in \{ab^k cd^k e \mid k \geq 0\} & \end{array}$$

Théorème. L_* est un point limite de la famille de langages $(L_k)_{k \geq 0}$ pour la classe des langages de prégroupe libre (simple) d'ordre 1/2.

Preuve. Les lemmes précédents montrent que $\forall k \geq 0, L_k \subsetneq L_{k+1}$ et $L = \bigcup_{k \in N} L_k$.

Corollaire. Les classes des grammaires k-valuées de prégroupe libre d'ordre $n/2, n \geq 1$ ne sont pas apprenables dans le modèle de Gold depuis des chaînes.

4 Conclusion

Nous avons montré que la classe des grammaires rigide de prégroupe libre n'est pas apprenable à partir des chaînes, même si on limite fortement l'ordre des types (ordre 1/2).

Un résultat de non apprenabilité pour la classe des prégroupes rigides avait déjà été présenté à (Foret, 2002), le point limite alors obtenu était une traduction d'un point limite pour le calcul de Lambek (Foret & Le Nir, 2002) qui est en fait à l'ordre 2. On aurait pu s'attendre à ce que l'ordre 1/2 soit apprenable. Notre résultat de non-apprenabilité peut donc surprendre mais il précise les contours de ce qui peut être entrepris ou non pour la mise au point d'algorithmes d'acquisition.

La non-apprenabilité à partir des chaines (sans structures) indique sans doute le besoin de structurer les exemples servant à l'apprentissage. C'est une voie de recherche qui nous intéresse tout particulièrement.

Références

V. ABRUSCI & C. CASADIO, Eds. (2001). *New Perspectives in Logic and Formal Linguisitics, Proceedings Vth ROMA Workshop*. Bulzoni Editore.

BARGELLI D. & LAMBEK J. (2001). An algebraic approach to french sentence structure. In (de Groote *et al.*, 2001).

BONATO R. & RETORÉ C. (september 2001). Learning rigid lambek grammars and minimalist grammars from structured sentences. *Third workshop on Learning Language in Logic, Strasbourg*.

BUSZKOWSKI W. (2001). Cut elimination for the lambek calculus of adjoints. In (Abrusci & Casadio, 2001).

CASADIO C. & LAMBEK J. (2001). An algebraic analysis of clitic pronouns in italian. In (de Groote *et al.*, 2001).

P. DE GROOTE, G. MORILL & C. RETORÉ, Eds. (2001). *Logical aspects of computational linguistics : 4th International Conference, LACL 2001, Le Croisic, France, June 2001*, volume 2099. Springer-Verlag.

DUDAU-SOFRONIE, TELLIER T. (décembre 2001). Learning categorial grammars from semantic types. In *13e Amsterdam Colloquium*, Palaiseau.

FORET A. (2002). Some unlearnability results for lambek categorial and pregroup grammars (unpublished). In *Gracq, ESSLII*, Trento, Italy.

FORET A. & LE NIR Y. (2002). Lambek rigid grammars are not learnable from strings. In *COLING'2002, 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

KANAZAWA M. (1998). *Learnable classes of categorial grammars*. Studies in Logic, Language and Information. FoLLI & CSLI. distributed by Cambridge University Press.

LAMBEK J. (1999). Type grammars revisited. In A. LECOMTE, F. LAMARCHE & G. PERRIER, Eds., *Logical aspects of computational linguistics : Second International Conference, LACL '97, Nancy, France, September 22–24, 1997 ; selected papers*, volume 1582 : Springer-Verlag.

LAMBEK J. (2001). Mathematics and the mind. In (Abrusci & Casadio, 2001).